LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-CONF-204182

Visualizations in Hostile Rapid Scan Forensics

Tony Bartoletti

May 17, 2004

DOE Computer Security Group Training Conference
Kansas City, MO, United States
May 24, 2004 through May 27, 2004

Visualizations in Hostile Rapid Scan Forensics

Tony Bartoletti

Computer Incident Advisory Capability

May, 2004

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

Topics to be covered

- What to do with All That Data …
- Rudimentary Visualizations
- Techbase-funded study to characterize services by traffic patterns
- LDRD-funded study to identify adversaries and their “true” network locations through packet-timing characteristics
- Packet TTL Distributions

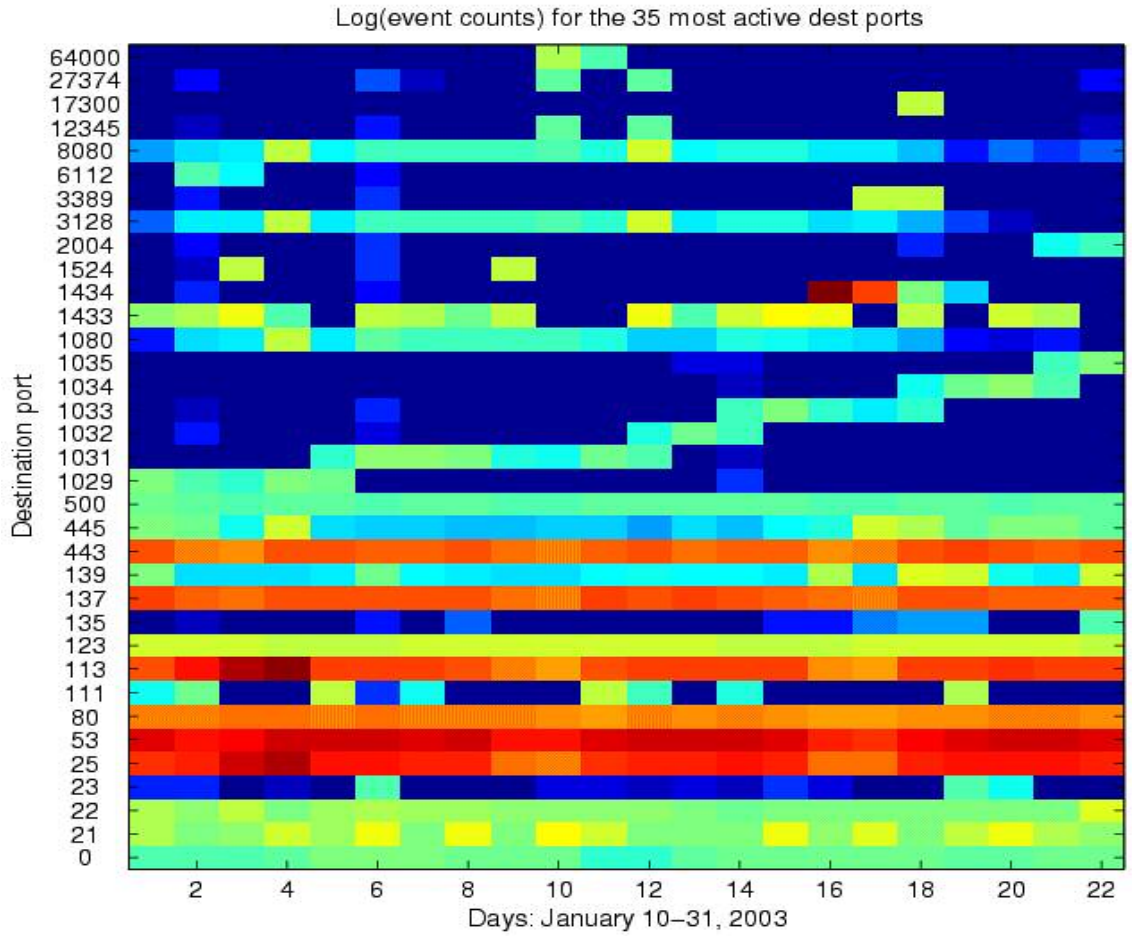
What to do with All That Data

- Network session and scan capture provides context for intrusion detection
- Known-Needle in haystack searches
- Broad trend and anomaly detection
- **Esoteric studies**

Exploratory methods reveal both normal and suspect patterns

Events: 4,252,862

Unique DestPorts: 20,809



~160,000 hits

~23,000 hits

~3,000 hits

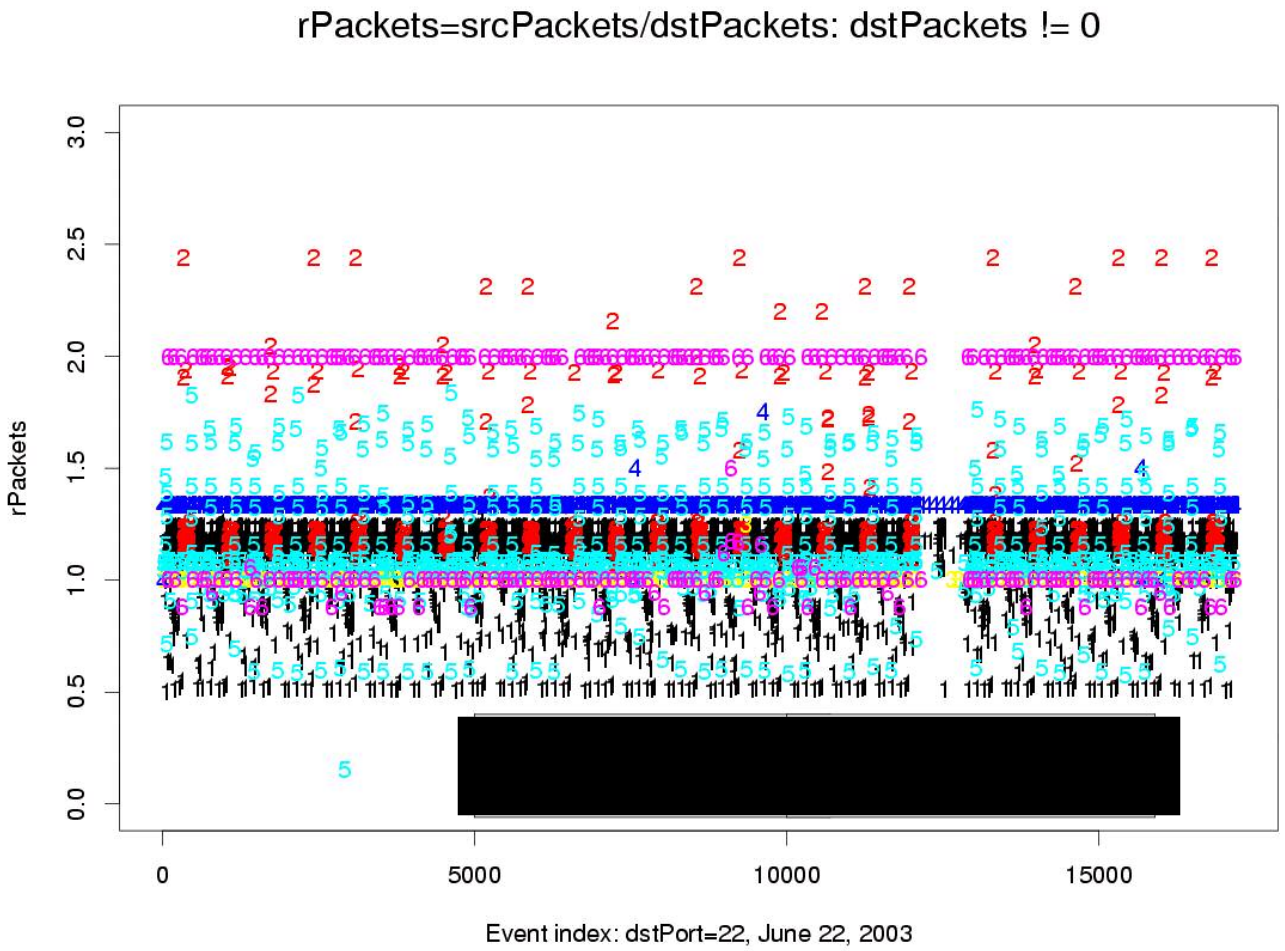
~400 hits

~55 hits

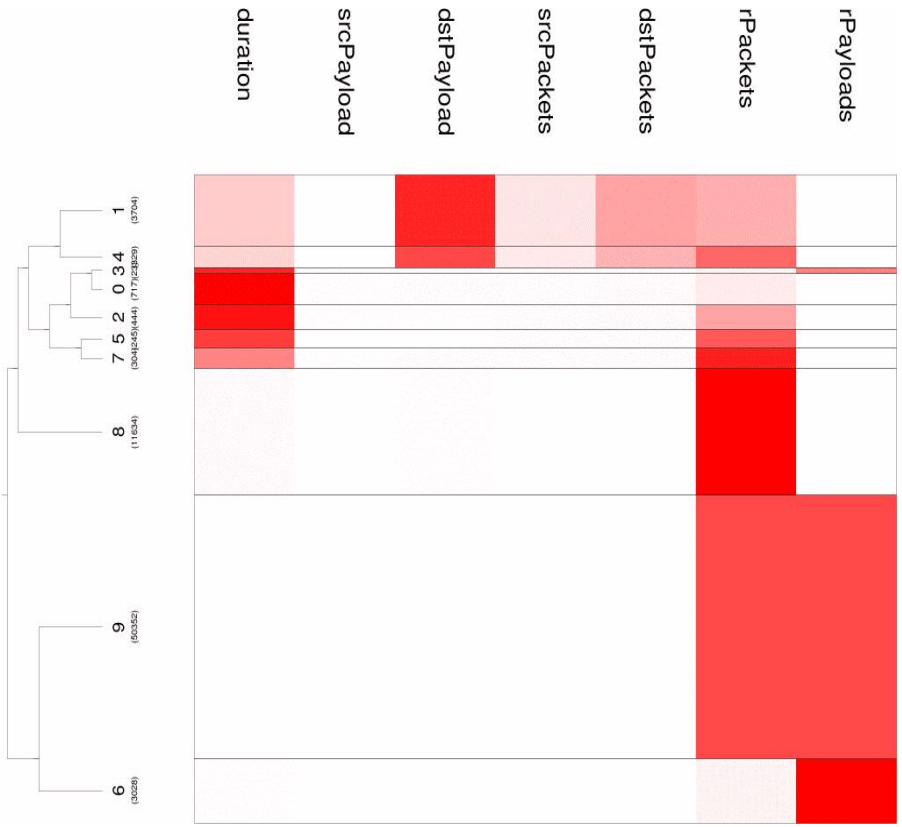
~7.5 hits

Packet-Ratio distributions reveal characteristic service patterns

Events: 15,475



Clustering partitions the data into groups: similar events belong to the same cluster



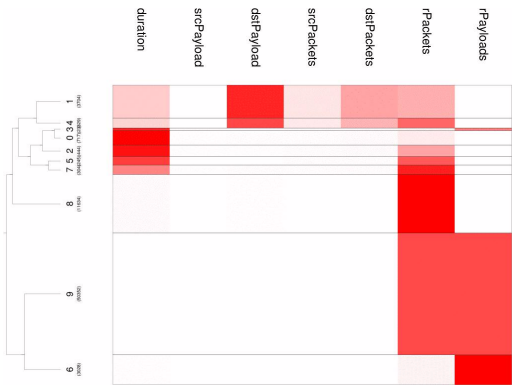
- 70,780 port22 events divided in ten clusters using CLUTO

- Event ***i*** is described by its 7D feature vector

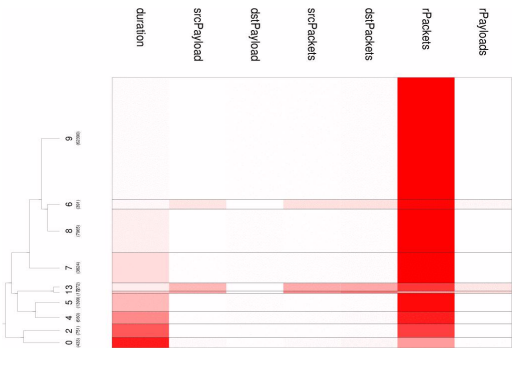
$$v_i = (duration, srcPayload, dstPayload, srcPackets, dstPackets, rPackets, rPayloads)$$

- Events ***i*** and ***j*** are in the same cluster if $dist(v_i, v_j)$ is "small"
- The rows show the clusters
  - ***The darker the color the larger the average value of the corresponding feature in that group***
  - ***The height of the cluster is proportional to the number of elements in that group***

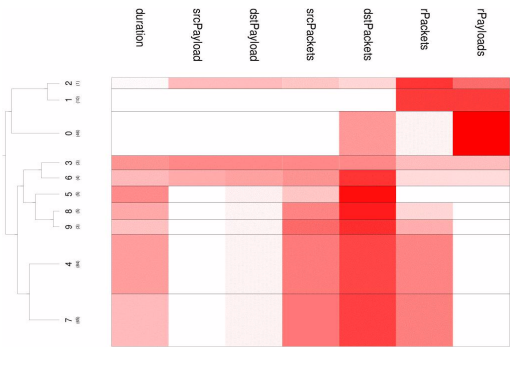
Different ports exhibited different clustering profiles



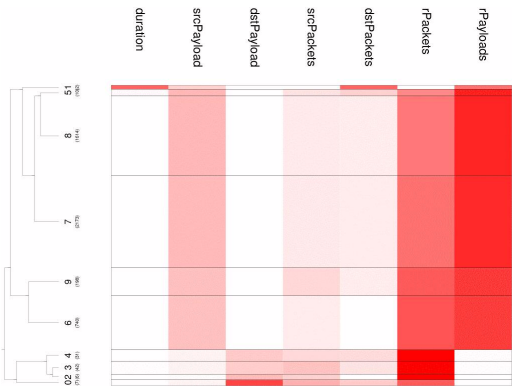
Port 22



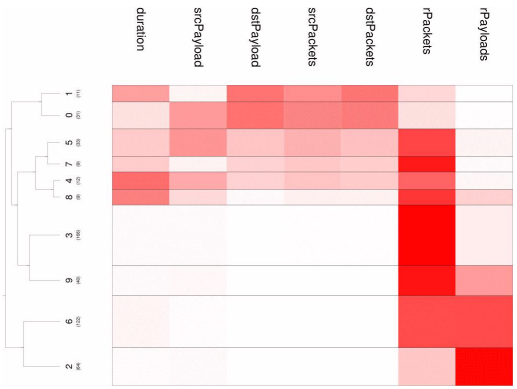
Port 25



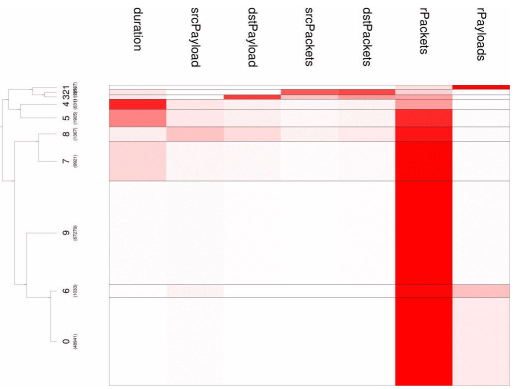
Port 135



Port 53



Port 8000

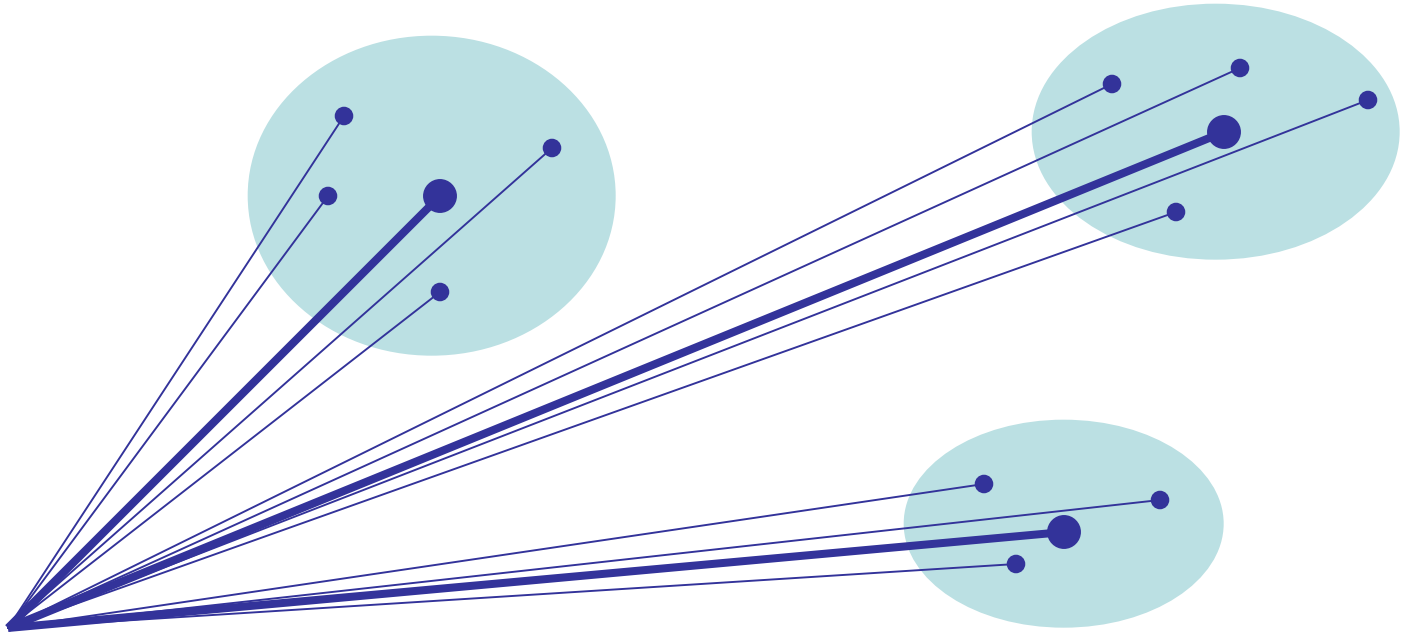


Port 80

FY04 Techbase-Funded Project

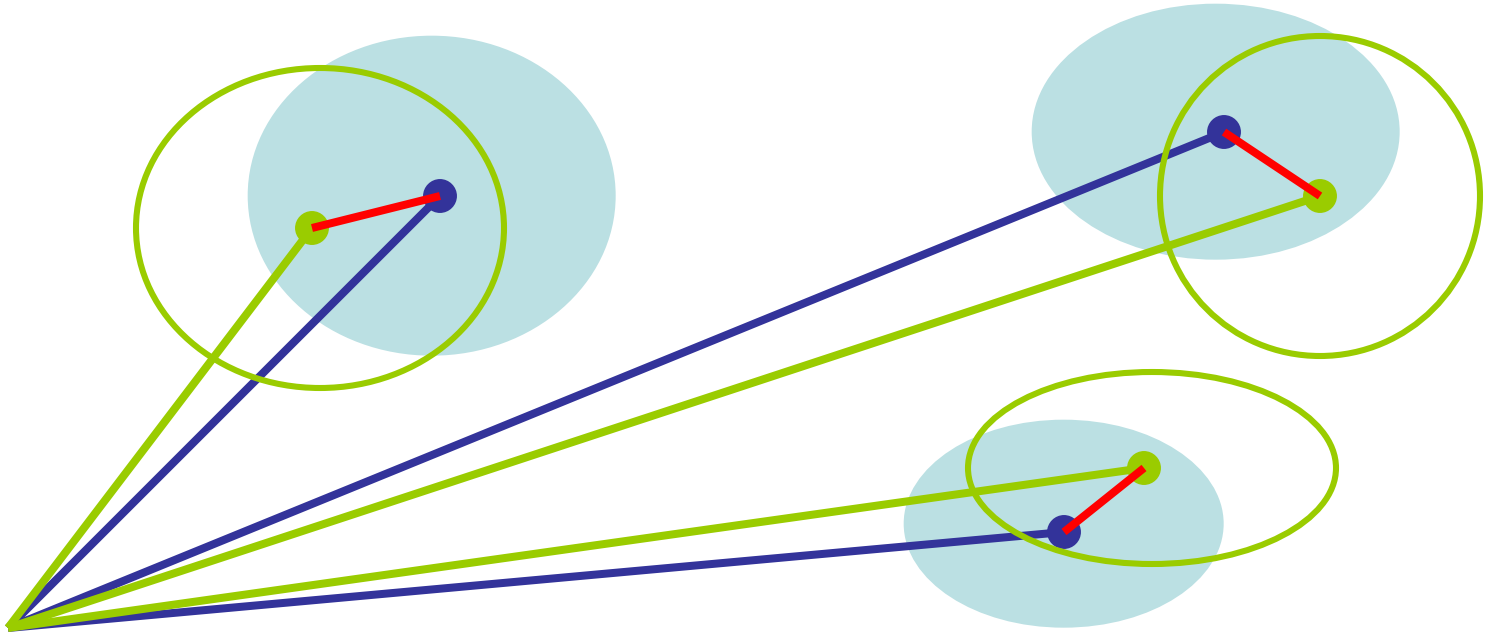
- Investigate Port/Service Session Clusters
- Develop metrics for distinguishing activity characteristic of, or anomalous to expectations, by port/service.
- Outline prototype system for operations, including performance expectations.

Use Cluster-Set Centroids to Characterize Service Behaviors



Port X Cluster Set and Cluster Centroids

Measure “Distance between” Cluster Sets via “Greedy Algorithm”



Sum “red” displacements as a measure of Cluster-Set Variation

An LDRD Proposal for
Exploratory Research in the Disciplines

Internet Ballistics: Identifying Internet Adversaries Despite Falsified Source Addressing

Tony Bartoletti, PI
Computer Incident Advisory Capability
April, 2004

Abstract

Visualizations in high-volume network attack traffic suggest attackers leave a “voiceprint” sufficient to support a degree of identification despite obfuscations in source IP address.

There is great value in determining the extent to which this form of identification is effective.

Every Day …

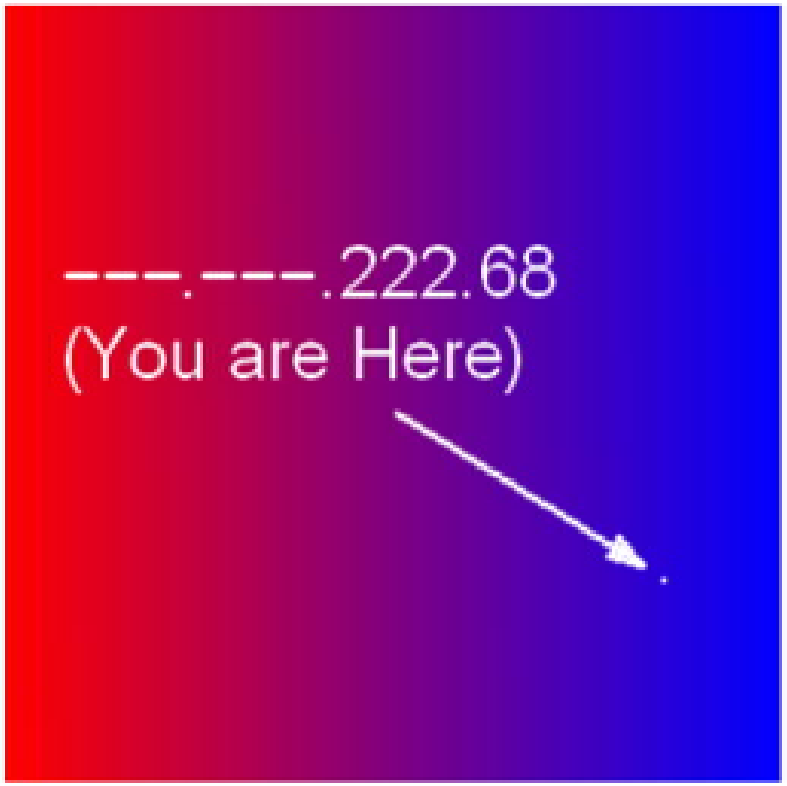
- Thousands of network attacks are launched against LLNL and other DOE/NNSA Sites
- Millions of hostile probes attempt to find and exploit weaknesses in computer services
- Roughly 25% of attackers probe more than half of a class-B subnet (64,000 addresses)
- Many attackers “visit” regularly, and many do not (if source IP address is the identity criteria)

Unexpected discovery in high volume packet traffic

Visualizations in packet arrival timing against target address space are often highly distinct and correspond to apparent source IP address

- Due to design of attacker's algorithms?
- Due to attacker's complement of running processes?
- Due to physical system constraints (memory, swap size, …)?
- Due to network location (nature of intervening routers)?

A “3D” graph of a complete class-B Scan



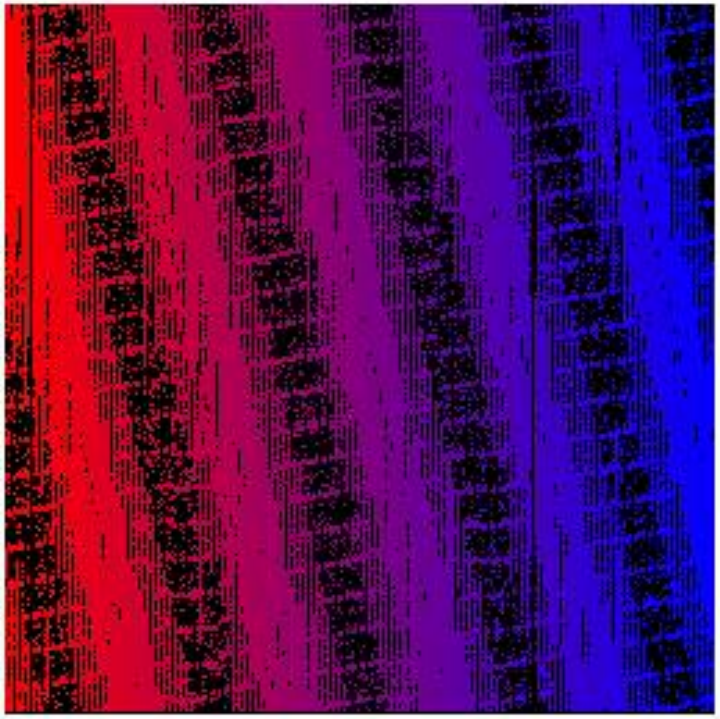

```
SrcIP: 0115
CB-Port:     3-443  (T)
Probes:      196325   Redun = 3.00
Dests:       65536    CB % = 100.00
ScanET:      2192     PPAS =  89.56
ActiveSecs:  2193     AS % = 100.00
ActiveSegs:  1
MaxGap:      0
IF-Factor:   2.68
```

Graph of probes to ---.---.X.Y: X,Y in [0,255] for single source and port.

Time, relative to scan start and end, is depicted by a red-blue gradient.

Scans as complete and uniform as this are rare.

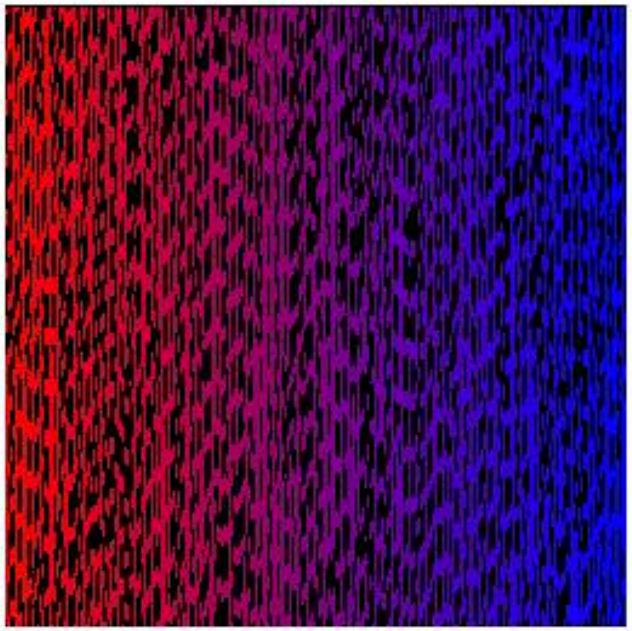
40% of scan-time graphs are highly distinct



SrcIP: 0284
CB-Port: 0002-80 (T)
Probes: 53025 Redun = 1.25
Dests: 42497 CB % = 64.85
ScanET: 497 PPAS = 106.69
ActiveSecs: 383 AS % = 76.91
ActiveSegs: 111
MaxGap: 2
IF-Factor: 1.50

Init_UTC_Offset: 1152226

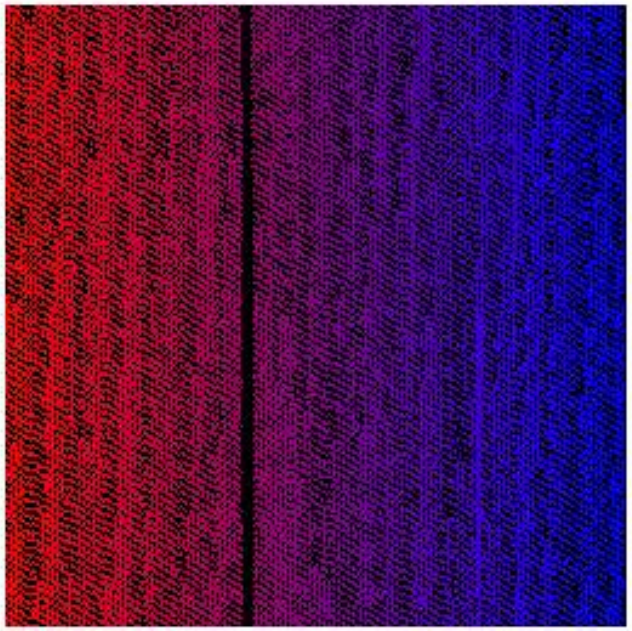
PLOT: First Probe



SrcIP: 0140
CB-Port: 0003-22 (T)
Probes: 35250 Redun = 1.00
Dests: 35225 CB % = 53.75
ScanET: 216 PPAS = 163.19
ActiveSecs: 146 AS % = 67.28
ActiveSegs: 71
MaxGap: 2
IF-Factor: 1.56

Init_UTC_Offset: 249433

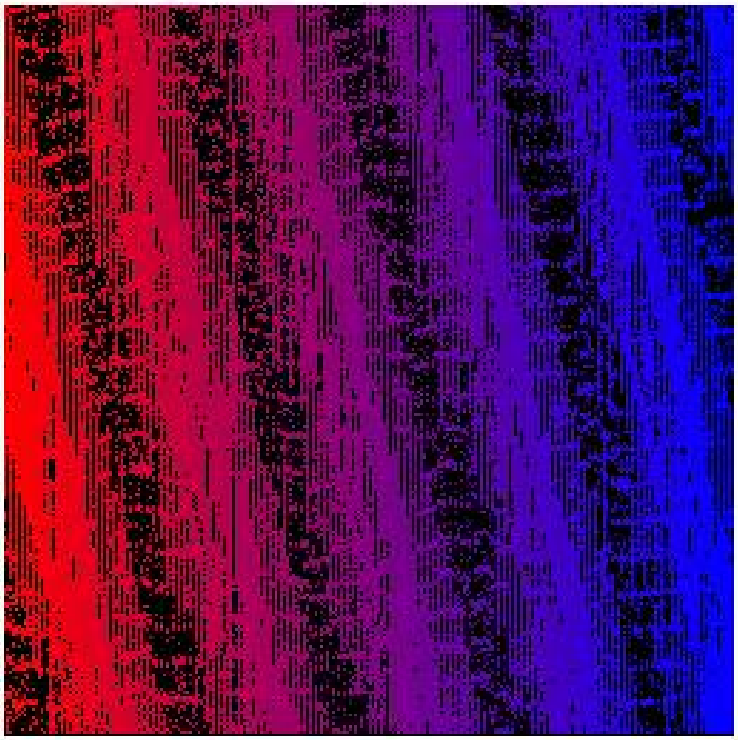
PLOT: First Probe



SrcIP: 0249
CB-Port: 0001-17300 (T)
Probes: 59675 Redun = 1.58
Dests: 37736 CB % = 57.58
ScanET: 2194 PPAS = 27.20
ActiveSecs: 2155 AS % = 98.18
ActiveSegs: 2
MaxGap: 40
IF-Factor: 18.42

Init_UTC_Offset: 589680

PLOT: First Probe



SrcIP: 0284

CB-Port:	0001-80	(T)
Probes:	53325	Redun = 1.25
Dests:	42589	CB % = 64.99
ScanET:	500	PPAS = 106.65
ActiveSecs:	371	AS % = 74.05
ActiveSegs:	128	
MaxGap:	2	
IF-Factor:	1.79	

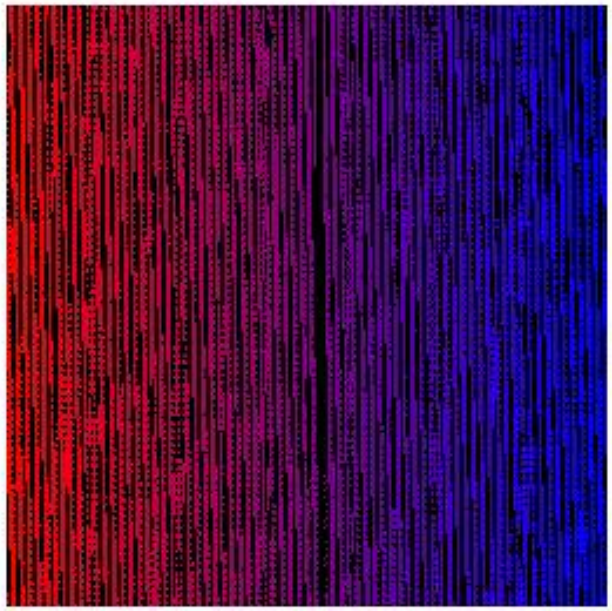
Init_UTC_Offset: 1061074

PLOT: First Probe

… Wait! Haven't we seen this one before? (see slide 17)

(Same apparent source, Different subnet, 25 hours apart)

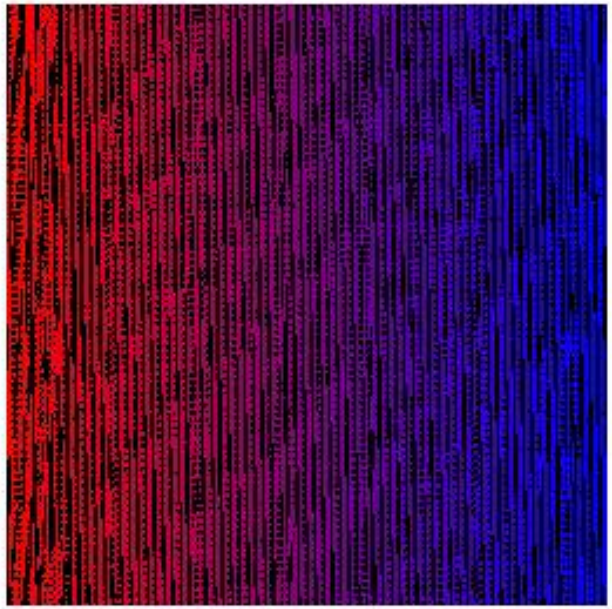
Additional samples on slides 20 and 21



SrcIP: 0111
CB-Port: 0001-21 (T)
Probes: 50850 Redun = 1.50
Dests: 33954 CB % = 51.81
ScanET: 1914 PPAS = 26.57
ActiveSecs: 1207 AS % = 63.03
ActiveSegs: 550
MaxGap: 22
IF-Factor: 14.44

Init_UTC_Offset: 45927

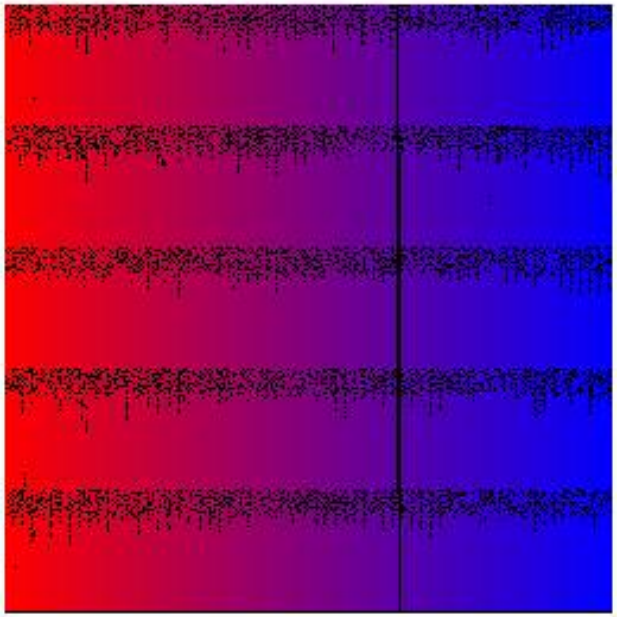
PLOT: First Probe



SrcIP: 0111
CB-Port: 0002-21 (T)
Probes: 52150 Redun = 1.51
Dests: 34558 CB % = 52.73
ScanET: 1933 PPAS = 26.98
ActiveSecs: 1293 AS % = 66.86
ActiveSegs: 518
MaxGap: 2
IF-Factor: 10.80

Init_UTC_Offset: 235430

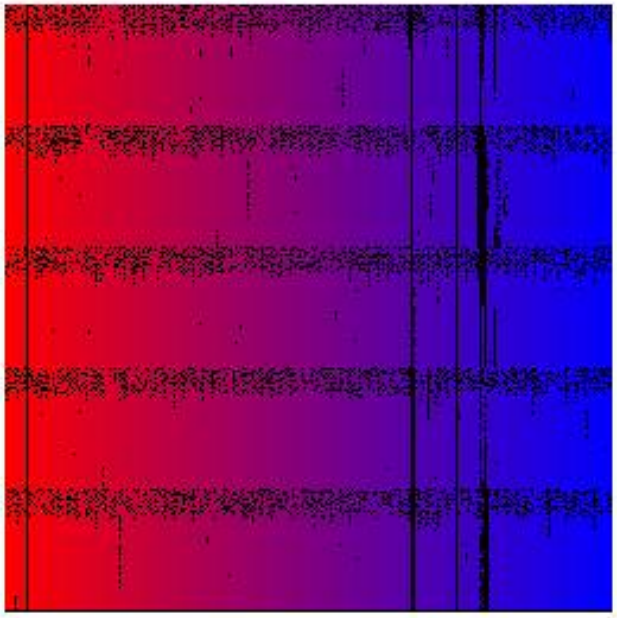
PLOT: First Probe



SrcIP: 0135
CB-Port: 0001-80 (T)
Probes: 81925 Redun = 1.39
Dests: 59011 CB % = 90.04
ScanET: 4500 PPAS = 18.21
ActiveSecs: 2007 AS % = 44.59
ActiveSegs: 1274
MaxGap: 27
IF-Factor: 8.16

Init_UTC_Offset: 602016

PLOT: First Probe

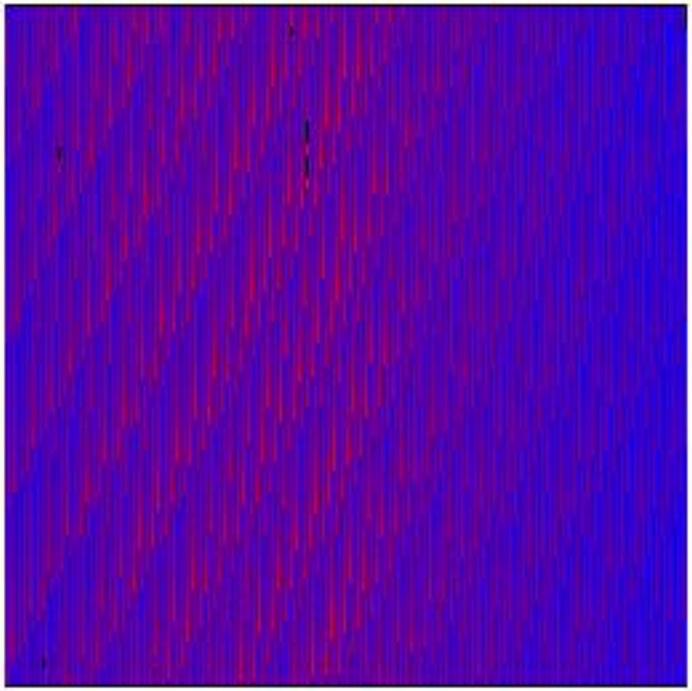


SrcIP: 0135
CB-Port: 0002-80 (T)
Probes: 81525 Redun = 1.40
Dests: 58428 CB % = 89.15
ScanET: 4497 PPAS = 18.13
ActiveSecs: 1975 AS % = 43.91
ActiveSegs: 1259
MaxGap: 27
IF-Factor: 5.86

Init_UTC_Offset: 1051916

PLOT: First Probe

Even “Smooth” scans have structure

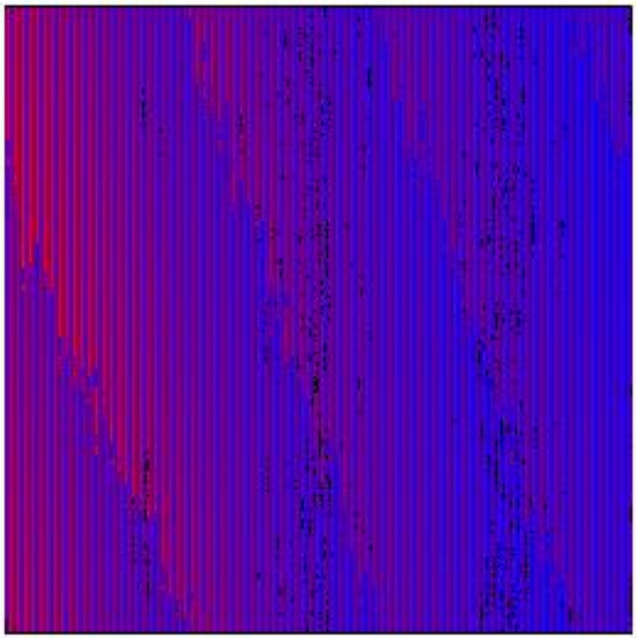


SrcIP: 0031
CB-Port: 0001-139 (T)
Probes: 64989 Redun = 1.00
Dests: 64989 CB % = 99.17
ScanET: 615 PPAS = 105.67
ActiveSecs: 616 AS % = 100.00
ActiveSegs: 1
MaxGap: 0
IF-Factor: 4.35

Init_UTC_Offset: 405027

PLOT: FirstProbe Rate Gradient

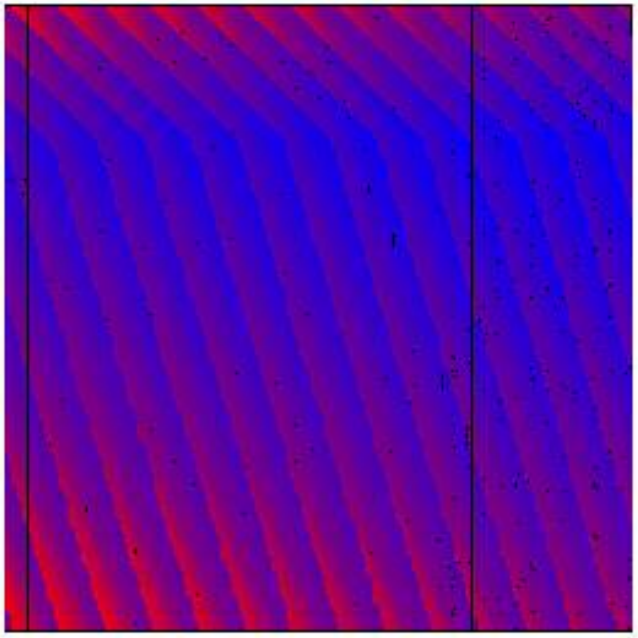
Differencing actual arrival times with that expected of a “perfect” gradient, or applying discrete derivatives, reveal finer structural features.



SrcIP: 0044
CB-Port: 0001-80 (T)
Probes: 64150 Redun = 1.00
Dests: 64150 CB % = 97.89
ScanET: 86 PPAS = 745.93
ActiveSecs: 87 AS % = 100.00
ActiveSegs: 1
MaxGap: 0
IF-Factor: 2.41

Init_UTC_Offset: 17554

PLOT: FirstProbe Rate Gradient



SrcIP: 0272
CB-Port: 0002-135 (T)
Probes: 64246 Redun = 1.00
Dests: 64246 CB % = 98.03
ScanET: 1549 PPAS = 41.48
ActiveSecs: 1540 AS % = 99.35
ActiveSegs: 3
MaxGap: 5
IF-Factor: 19.43

Init_UTC_Offset: 1114081

PLOT: FirstProbe Rate Gradient

Why Care? Value of consistent adversary identification

Critical to cyber security and Internet counterintelligence

- "Major Players" may serve as early warning of new exploits
- Adversary hierarchies and alliances can be mapped
- Adversary correlation is critical to damage assessment

Fundamental problem in Internet source identification

Source IP address as primary identification is unreliable

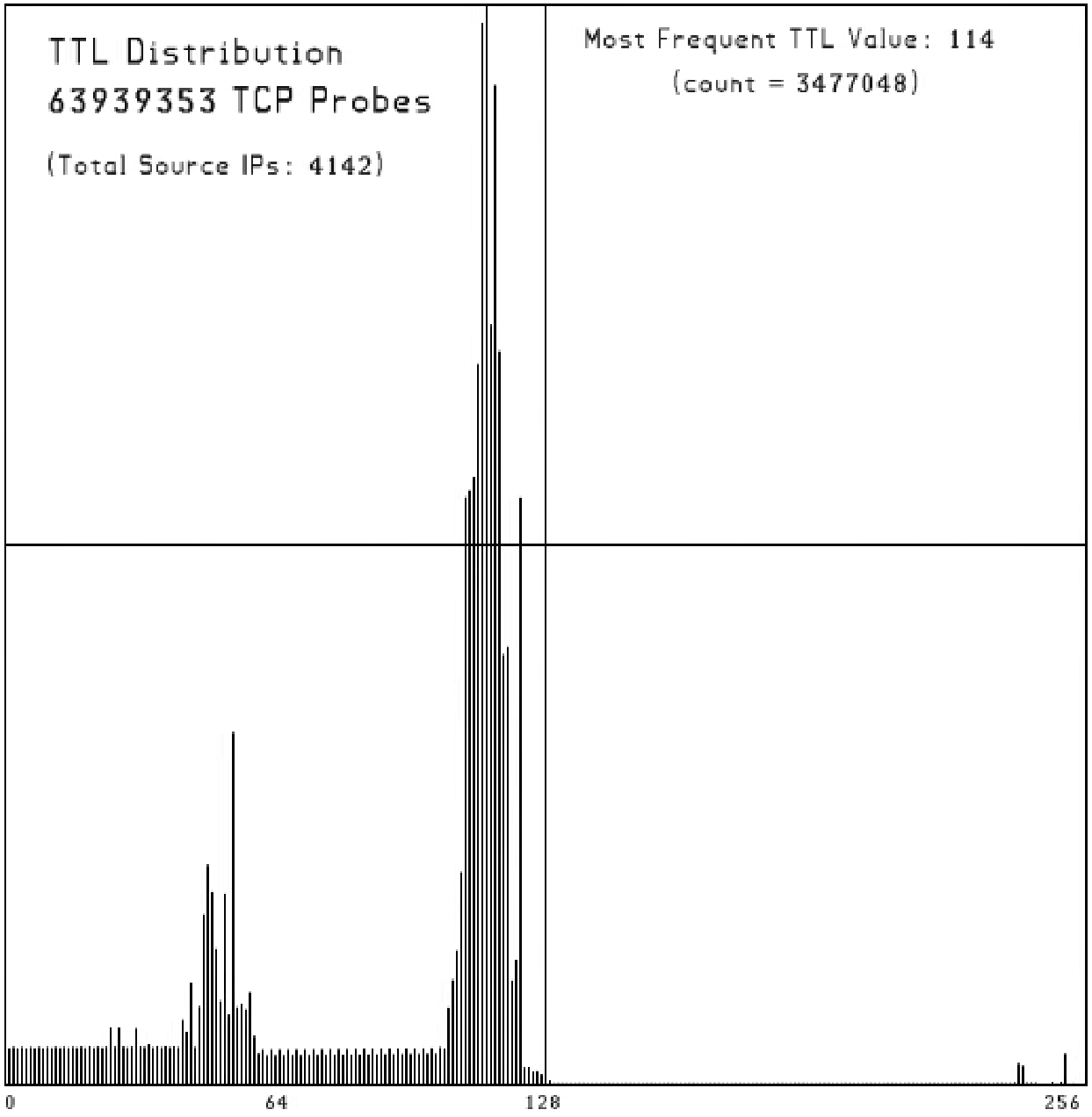
- IP Addresses are easily forged
- Misappropriated systems are often employed
- ISPs apply dynamic addressing and address translation

Research

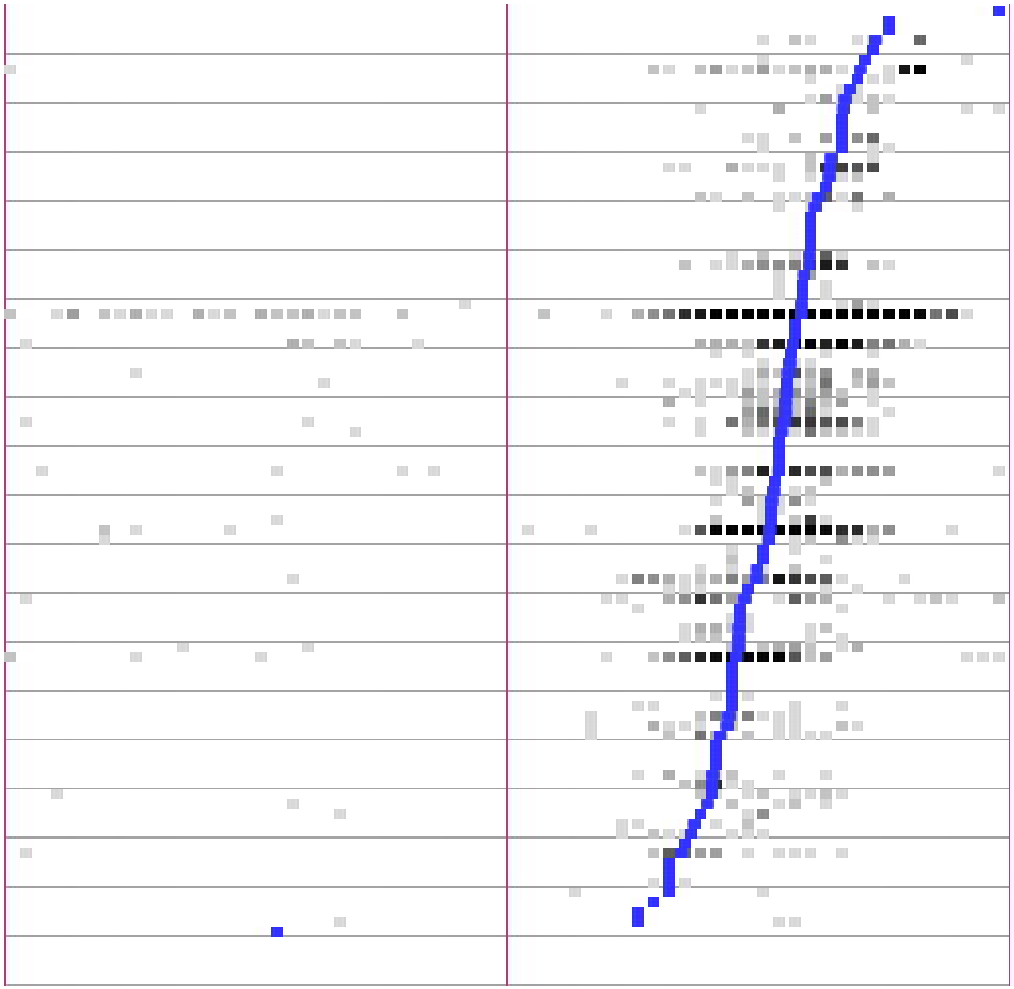
- Reduce arrival-time data to frequency spectrum vectors using wavelet analysis
- Attempt to identify vector components sensitive to source or network variations independently
- Develop a metric space suitable for seeking “closeness” in vectors/components
- Provide a foundation for confidence intervals in attribution hypotheses
- (Need tools like TIPS for reliable data capture)

A Study of Packet TTL Distributions

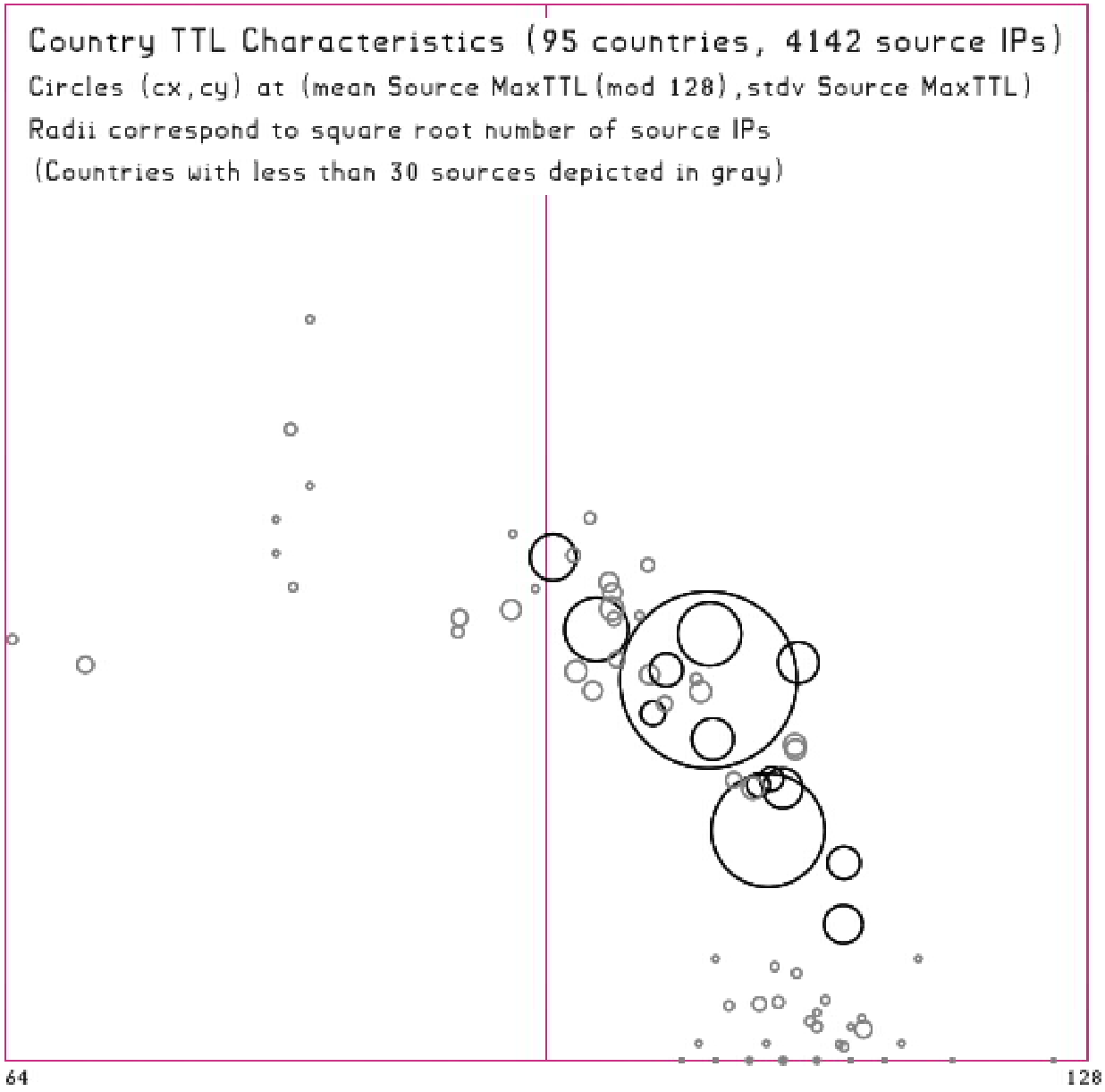
- Distribution of TTL values seen in 64 million TCP SYN probes
- Distribution of Max TTL values per source IP, by country (4142 sources)
- Another visualization of TTL variations by country
- Per-Source TTL characterization.
- (Need tools like TIPS for reliable data capture)

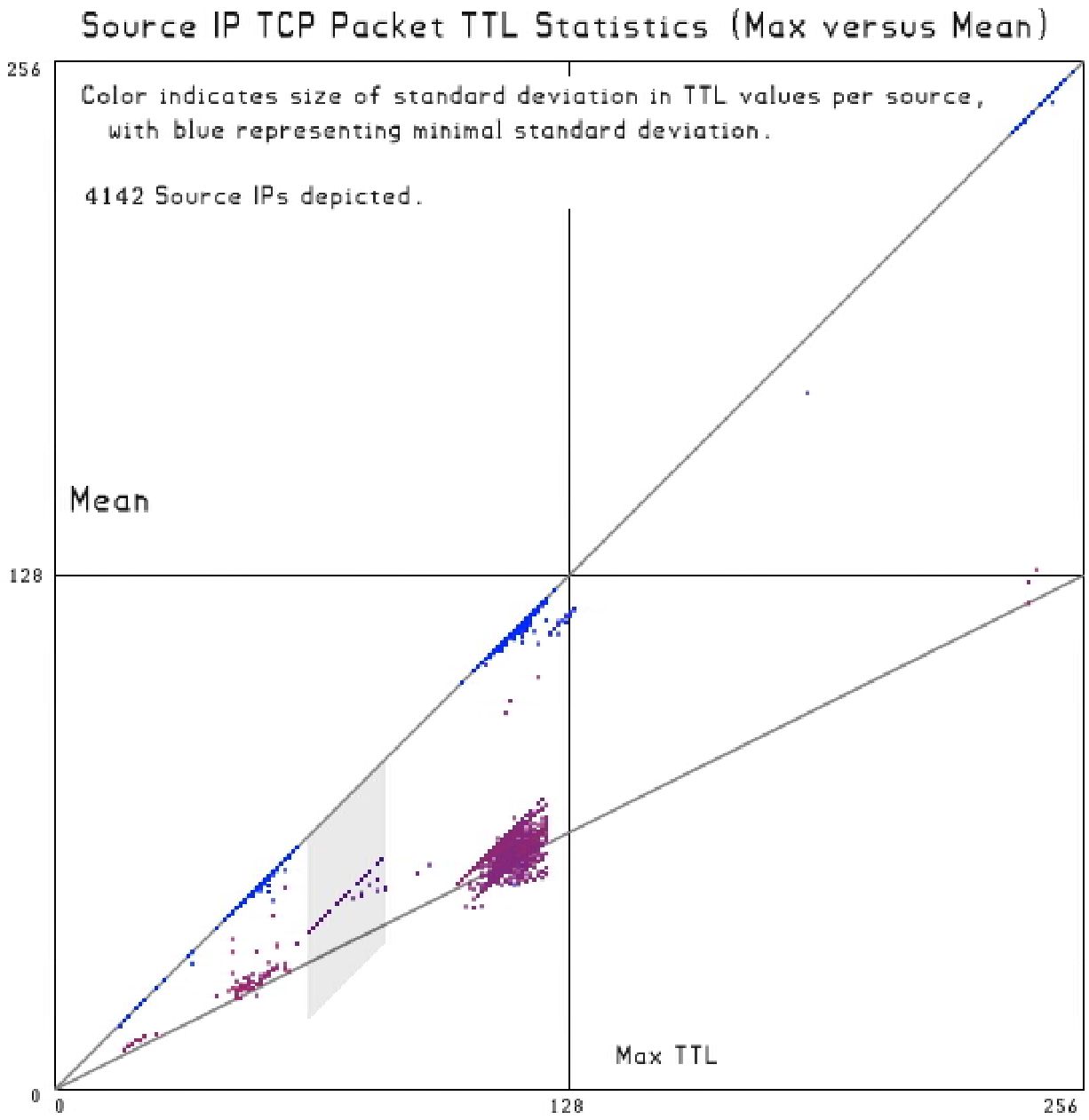

TTL Distribution
63939353 TCP Probes
(Total Source IPs: 4142)
Most Frequent TTL Value: 114
(count = 3477048)
0
64
128
256

Country TTL Characteristics (95 countries, 4142 source IPs)
Distribution of Source IP Max-TTL values taken modulo 64
(Countries sorted according to Mean of SourceMaxTTL)



SN PR CL KW GH TR AE NZ IN DM FI LT IT AT IE KZ NO DO SI
MO DE BH DK BO FR US -- NL KH CR HU TN CN RO MA PE BR LY
VN PA MN JO EC LK OM HK JP TW BE EG IR LV SE TT CO BY HR
MY MK CH JM ID DZ PS IL TH VE KR GB PL MT RU AR CZ LU UY
QA SG YU AU SK NG CA MX AG IS PH GR SA NI ES PT ZA UA EE


Country TTL Characteristics (95 countries, 4142 source IPs)
Circles (cx,cy) at (mean Source MaxTTL (mod 128), stdv Source MaxTTL)
Radii correspond to square root number of source IPs
(Countries with less than 30 sources depicted in gray)
64
128


Source IP TCP Packet TTL Statistics (Max versus Mean)
Color indicates size of standard deviation in TTL values per source,
with blue representing minimal standard deviation.
4142 Source IPs depicted.
Mean
Max TTL
256
128
0
0
128
256

Qualifications of the Principal Investigator

- *MS Mathematics, Oregon State University, 1987*
- *Tony Bartoletti has been a member of the DOE Computer Incident Advisory Capability (CIAC) since 1991.*
- *He has managed several security application tool development efforts such as the Security Profile Inspector for Networks, and Safepatch, receiving a Government Technology Leadership Award in 2000 for the latter.*
- *He is currently responsible for research methods and their applications under the CIAC data analysis regime.*

QUESTIONS?

- Tony Bartoletti (azb@llnl.gov)